文献–作者二分网络中基于路径组合的合著关系预测研究*

张金柱[1] 王小梅[2] 韩 涛[2]
[1](南京理工大学经济管理学院 南京 210094)
[2](中国科学院文献情报中心 北京 100190)

摘要:【目的】降低文献–作者二分网络在投影为合著网络过程中的信息丢失影响,形成适应特定二分网络的合著关系预测指标和方法,提高预测准确率和结果可解释性。【方法】首先构建文献–作者二分网络及其投影合著网络;接着抽取二分网络中的二阶路径和三阶路径表示作者间的关联关系;最后利用逻辑回归方法学习不同路径对于合著关系预测的贡献,由此形成文献–作者二分网络中基于路径组合的合著关系预测指标。【结果】在图书情报领域的实验证实,文献–作者二分网络在投影为合著网络过程中存在较大的信息丢失,并以合著关系预测准确率变化进行定量计算;逻辑回归方法适合学习不同路径对于合著关系预测的贡献,由此形成的路径组合指标准确率远远高出其他指标,并且预测结果更易解释。【局限】其他的多阶路径尚未引入到该模型中,方法通用性还需在其他领域进行验证。【结论】合著关系预测应直接在文献–作者二分网络上进行,以降低投影过程中的信息丢失影响;文献–作者二分网络上的路径组合指标是合著关系预测的最优指标;该方法可扩展应用到其他类型的二分网络中,如专利–发明人二分网络。

关键词: 文献–作者二分网络 路径组合指标 图书情报 合著网络 合著关系预测
分类号: G350

1 引 言

多学科交叉融会的大背景以及科研人员研究方向的专业化和精细化使得越来越多的科学研究由个人独立完成转变为科研团队协作完成,从而提高科研水平和科研效率。这就使得适应时代要求和特定主题的科研团队组建研究逐渐受到重视并引起了广泛关注。合著关系作为科研合作的重要体现,也是发现科研合作的重要途径[1],因此,作者合著可能性可以在一定程度上代表作者的科研合作可能性,进而为科研团队人员选择和搭配提供建议和参考[2]。

当前,合著关系预测主要在合著网络中进行,它以作者为节点,以合著关系为边,由于其节点和连边均为单一类型,因此属于单分网络的一种表现形式。合著网络中的合著关系预测就是尚未产生连边的节点对之间产生连边的可能性预测[3],应用和改进复杂网络中节点间的多种相关性计算指标,可以计算当前尚未产生合著关系的作者对的相关程度,并以相关程度表示作者对在未来产生合著的可能性[4]。作者对的相关性计算指标可以分为共同邻居及其改进指标、到达路径指标和随机游走指标[5],并已在多个领域中进行实验以比较不同指标的优劣,寻找合著关系预测的最优指标[6-7]。而合著网络是由文献–作者二分网络投影形成,投影过程中文献信息的丢失使得合著关系具体

通讯作者: 张金柱, ORCID: 0000-0001-7581-1850, E-mail: zhangjinzhu@njust.edu.cn。

*本文系国家自然科学基金青年基金"基于被引科学知识突变的突破性创新动态识别及其形成机理研究"(项目编号: 71503125)、教育部人文社会科学研究青年基金"异构知识网络中主题突变动态识别研究"(项目编号: 14YJC870025)和中央高校基本科研业务专项资金"基于专利引用科学知识突变的突破性创新动态识别方法与形成机理研究"(项目编号: 30915013101)的研究成果之一。

发生在哪些文献上难以跟踪，并可能导致合著关系预测的准确率降低[8-9]，因此，需要计算同一指标在二分网络及其投影合著网络上的准确率变化，量化表示信息丢失及其对合著关系预测的影响。这使得直接在二分网络上对合著关系进行预测成为一种新思路。二分网络由两种类型的节点构成，并且边只在不同类型的节点间存在，由此形成了相应的中心性指标、集聚系数、社团结构和演化模型[10]。二分网络上的关系预测主要是对单分网络上的指标在二分网络上进行映射，形成了共同邻居、局部路径等指标在二分网络上的对应表示[11]，并在商品–消费者、RNA–蛋白质和图书–借阅者等二分网络上应用，取得了相对较好的效果。然而，文献–作者二分网络上作者间的关联关系相对于合著网络更加多样和复杂，如何抽取和表示多种关联关系并明晰它们对于合著关系预测的贡献还需深入研究，如何融合多种关联关系形成合著关系预测的最佳指标还需进一步加强。

本文直接在文献–作者二分网络中抽取多种路径表示作者间的关联关系，并通过逻辑回归的机器学习方法学习不同路径对于合著关系预测的贡献，以学习到的权重系数组合多种路径形成二分网络中基于多路径组合的合著关系预测指标；在此基础上，对文献–作者二分网络及其投影合著网络的相关预测指标进行比较和分析，并通过准确率变化定量计算投影过程中的信息丢失。

2　文献–作者二分网络中的合著关系预测模型

文献–作者二分网络中合著关系预测模型包括三个部分，即：二分网络及其投影合著网络的构建、作者关联关系在二分网络中的路径表示及其组合、合著关系预测指标的评测。首先，设计二分网络投影为合著网络的方案，使得二分网络和投影网络在合著关系预测上具有高一致性，进而可以进行公平比较；接着在二分网络上抽取多种路径表示作者间的关联关系，作为合著关系发生的驱动因素，并使用机器学习的方法构建多路径组合指标；最后使用链路预测的方法对二分网络上的预测指标进行评测。

2.1　文献–作者二分网络及其投影网络构建

二分网络在投影为合著网络时存在信息丢失[8-9]，为了量化信息丢失，需要使得二分网络和投影网络在进行合著关系预测指标比较时具有高一致性。如图 1 所示，图 1 (a)表示文献–作者二分网络，而图 1 (b)为投影形成的对应合著网络，其中 P_i 为文献，A_i 为作者。

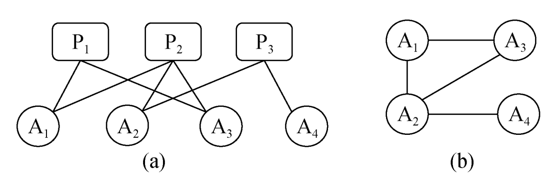


图 1　二分网络及其投影的合著网络

本文使用如下方法构建文献–作者二分网络及其对应的投影网络，并形成相应的训练集和测试集，为模型训练和结果评价提供数据基础。首先在二分网络中抽取所有的合著关系，并使用“作者–文献–作者”表示，如在图 1(a)的二分网络中，所有的合著关系表示为(A_1, A_3) : $[A_1P_1A_3, A_1P_2A_3]$, (A_1, A_2) : $[A_1P_2A_2]$, (A_2, A_3) : $[A_2P_2A_3]$和(A_2, A_4) : $[A_2P_3A_4]$。接着，依据 10 折交叉验证(10-Fold Cross Validation)方法得到训练集和测试集[7,12]，即：将数据集等分成 10 组，每组中的合著关系均从原数据集中随机抽取并且不重复，依次将每组数据作为一次测试集，余下的 9 组数据共同作为训练集，由此得到 10 组训练集和测试集。最后，使用合著关系对应的“作者–文献–作者”关系形成对应二分网络的训练集和测试集，如图 1 中，假设以(A_1, A_3)、(A_1, A_2)和(A_2, A_3)作为训练集，以(A_2, A_4)表示测试集，那么在二分网络中则是以$[A_1P_1A_3, A_1P_2A_3]$、$[A_1P_2A_2]$和$[A_2P_2A_3]$为训练集，以$[A_2P_3A_4]$为测试集。

这种训练集和测试集分割方法确保了二分网络与投影网络在进行合著关系预测指标比较时具有高一致性，然而，由于投影网络没有存储文献信息，使得两种网络仍存在一定的不一致性，这种不一致性正验证了投影过程中信息丢失的存在，并使得同一指标在二分网络和合著网络上的计算结果不同。举例来说，如果选择$[(A_1, A_3), (A_2, A_3), (A_2, A_4)]$作为训练集，$[(A_1, A_2)]$作为测试集，那么在二分网络训练集中对应的“作者–文献–作者”关系为$[A_1P_1A_3, A_1P_2A_3, A_2P_2A_3, A_2P_3A_4]$，而该训练集中的关系$[A_1P_2A_3, A_2P_2A_3]$会直接导致关系$A_1P_2A_2$发生，并不需要进行任何预测。

2.2　基于逻辑回归的多路径组合指标构建

二分网络构建后，需要从中抽取作者间的多种到

达路径表示作者间的关联关系，并基于逻辑回归的机器学习方法学习不同路径对于合著关系预测的影响和贡献，由此形成文献–作者二分网络中基于路径组合的合著关系预测指标。

(1) 二分网络中的路径表示和提取

文献–作者二分网络中，作者间的关联关系是通过文献形成的，如作者间的合著关系可以通过"作者–文献–作者"表示；两个作者的共同邻居数目可以通过"作者–文献–作者–文献–作者"(APAPA)的路径数目表示，单以作者来说，两个作者间的到达路径长度为 2，并与合著网络中作者间的共同邻居相对应，因此称该种关联关系为二阶路径；两个作者的合著者产生的合著关系可以通过"作者–文献–作者–文献–作者–文献–作者"(APAPAPA)路径表示，单以作者来说，两个作者间的到达路径长度为 3，因此称该种关联关系为三阶路径。在图 1 中，$A_1P_1A_3$ 表示作者 A_1 和 A_3 具有合著关系；$A_1P_2A_2P_3A_4$ 表示作者 A_2 是作者 A_1 和 A_4 的共同邻居，由于 A_1 和 A_4 间只有一条类似路径，因此 A_1 和 A_4 的二阶路径数目为 1；$A_3P_1A_1P_2A_2P_3A_4$ 表示作者 A_3 的合著者 A_1 与 A_4 的合著者 A_2 具有合著关系，由于 A_3 和 A_4 间只有一条类似路径，因此 A_3 和 A_4 的三阶路径数目为 1。

本文中以文献–作者二分网络中的二阶路径和三阶路径表示作者间的关联关系，并且由二阶路径和三阶路径可以扩展形成四阶路径和更高阶路径。

(2) 基于逻辑回归的多路径组合方式

二分网络中的多种路径均可能对合著关系预测产生影响，而每种路径的贡献可能并不相同，因此，需要使用机器学习的方法在训练集中学习每种路径的权重系数表示该路径对于合著关系预测的贡献，进而形成基于多路径组合的合著关系预测指标。

逻辑回归(Logistic Regression)是机器学习中的一种分类模型，在数据挖掘、疾病自动诊断和经济预测等领域均有较多应用[13]。逻辑回归常用来解决二分类问题，它基于一个或多个自变量(即二分类的影响因素)来计算该数据属于二分类中特定类别的概率，通过在训练集中学习到的不同影响因素的权重系数，便可以预测新数据的所属分类并计算属于特定类别的概率。

应用到多路径组合指标的构建时，自变量是二阶路径数目和三阶路径数目，而因变量是合著关系是否发生，发生即为 1，没有发生即为 0。对于训练集中的每一对合著关系(i, j)，X_k 为二维向量，用来存储二阶路径和三阶路径的数目，y_k 则表示合著关系是否发生。举例来说，当作者 i 到 j 的二阶路径数目为 2、三阶路径数目为 6 时，合著关系发生，那么 $X_k=[2, 12]$, $y_k=1$。逻辑回归方法则使用 10 折交叉验证中的训练集作为正例(Posive)，并随机抽取同样数量的负例(Negative)一起作为训练集，进而应用 Python 语言的 scikit-learn 机器学习工具包实现逻辑回归，对应的类为"sklearn.linear_model.LogisticRegression"。通过输入训练集中的多个 X_k 和 y_k，得到二级路径和三阶路径的权重，进而形成多路径组合指标，并以此计算尚未产生合著关系的作者对之间合著的概率，对合著关系进行预测。

2.3 基于链路预测的指标评测

链路预测经常被用来定量评测复杂网络上的相关性指标优劣[7]，而文献–作者二分网络及其投影网络均属于复杂网络，并且合著关系预测指标也是相关性指标的一种，因此，可以应用链路预测的理论和方法对多种预测指标进行评测。定义 G=(V, E)为文献–作者二分网络，其中 V 为作者集合，E 为"作者–文献–作者"表示的合著关系集合，该网络中的合著关系全集 U 的个数为 N×(N–1)/2。给定一种合著关系预测指标，计算尚未产生合著关系的作者对(x, y)∈(U–E)的合著可能性，并按照可能性从大到小排序，排在最前面的作者对将来合著的可能性最大。

为了评价合著关系预测指标，将合著关系集合 E 通过 10 折交叉验证方式分为 10 组训练集 E^T 和测试集 E^P，在训练集上利用合著关系预测指标计算作者对的合著可能性，并在测试集上对计算结果的准确性进行评价。链路预测中衡量准确性的指标主要包括 AUC(Area Under Roc Curve)和 Precision(准确率)[7]，其值均为 10 次计算的平均结果。AUC 和准确率对指标精确度的衡量侧重点不同。AUC 从整体上衡量指标的精确度，AUC 值的区分度较低，即多个指标的 AUC 值差异较小，使得预测准确率较低的指标其 AUC 值可能仍然较大；Precision 则衡量排在前 L 位的合著关系是否预测准确，L 的取值可以自由确定，本文中使用 R-Precision 对合著关系预测准确率进行评价，既考虑准确性，同时考虑合著关系预测结果的排序，此时 L 为测试集中的合著关系数目。

3　实证分析

以图书情报领域的数据为例，构建文献–作者二分网络和投影形成的对应合著网络，在二分网络和合著网络上应用合著关系预测指标，计算预测准确率和 AUC 值，从而发现文献–作者二分网络在投影为合著网络时存在多少信息丢失、计算二分网络中不同路径对于合著关系预测的贡献、验证基于路径组合的指标和方法有效性。

3.1　数据说明

从 WoS(Web of Science)上下载被 SCIE (Science Citation Index Expanded)收录的学科分类为图书情报(Information Science & Library Science)的相关数据，对应时间段为 2005 年到 2009 年。同时，去除了 *Scientist* 期刊的相关数据，原因在于该期刊包含的论文数量众多并且论文长度很短，并且该期刊同时属于其他多个学科分类。如果包含该期刊的数据，会导致频次较高的作者均为该期刊上发表论文的作者，使得实验结果的可信度降低。所用数据集对应的检索表达式为:

(WC = Information Science & Library Science)
AND LANGUAGE: (English)
AND DOCUMENT TYPES: (Article)
Indexes=SCI-EXPANDED
Timespan=2005-2009
Refined by: [excluding] SOURCE TITLES: (Scientist)

数据预处理过程主要是删除匿名作者信息，即删除掉作者名为“[anonymous]”的相关作者。在此基础上，选取出现频次大于或等于 3 的作者及其对应文献构建文献–作者二分网络，并投影形成对应的合著网络，相关数据说明如表 1 所示。其中，孤立作者数目是指在选取出的高频作者中没有产生合作的作者数目，训练集中的合著关系数占总数的 90%，测试集中的合著关系数占 10%。

表 1　数据说明

时间段	作者数目	孤立作者数目	合著关系总数	训练集中合著关系数	测试集中合著关系数
2005-2009	911	159	1 183	1 064	119

3.2　文献–作者二分网络投影过程中的信息丢失

文献–作者二分网络在投影为合著网络的过程中，存在较大的信息丢失。CN(Common Neighbor)和二阶路径指标分别表示合著网络和对应二分网络中的共同邻居数目，在合著网络中，CN 指标的正确率和 AUC 分别为 27.1%和 85.4%，而在文献–作者二分网络中，二阶路径指标的正确率和 AUC 分别为 48.3%和 85.5%。其中，二分网络的正确率较合著网络高出了 21.2%，提高了 78.2%，定量表示了文献–作者二分网络在投影为合著网络过程中的信息丢失；与此同时，AUC 作为一个宏观评价指标，几乎没有变化，区分度很小。以上结果表明，合著关系预测的准确率降低与二分网络投影为合著网络的信息丢失密切相关，并可通过准确率变化定量表示信息丢失的多少，为投影过程中的信息丢失定量计算提供了一种新思路。

投影形成的合著网络无法体现文献信息导致合著关系预测的准确率降低，同时使得合著关系预测结果的解释难度加大。如在第一次实验成功预测的前 10 对合著关系中，CN 和二阶路径均成功预测 Bates DW 和 Jenter CA 会产生合著关系，CN 的数目为 6，而二阶路径(APAPA)数目为 14 且都表示共同邻居；另一方面，在前 119 对(与测试集中的数目相等)合著关系中，二阶路径成功预测 Markpin T 和 Sombatsompop N 产生合著关系，而 CN 没有预测出，此时 CN 的数目为 2、二阶路径(APAPA)数目为 13。这些都说明部分共同邻居关系随着投影过程也出现了丢失，最终导致了投影网络中合著关系预测的准确率降低。另一方面，文献–作者二分网络更易于跟踪作者对在哪些文献上进行合著，进而发现合著关系发生的原因和动机，更适合对合著关系预测进行解释和说明。

综上，文献–作者二分网络在投影为合著网络过程中存在较大的信息丢失，使得合著关系预测准确率大幅降低，因此，合著关系预测应直接在文献–作者二分网络上进行，以降低在投影为合著网络过程中的信息丢失影响，同时增加结果的可解释性。

3.3　路径组合指标与其他指标的结果比较分析

本文选取二分网络上的三种路径指标进行比较分析，分别是：与共同邻居(Common Neighbor)对应的二阶路径指标(二阶路径的数目)；与局部路径指标(Local Path)对应的路径组合指标，均表示二阶路径和三阶路径的组合，而局部路径指标 1 表示三阶路径的权重固定为 0.1，局部路径指标 2 表示三阶路径的权重固定为 0.01；以及表示三阶路径数目的三阶路径指标。为了对

合著关系预测指标进行公平全面比较，本文选取合著网络中的共同邻居和资源分配指标作为共同邻居及其改进指标的代表、局部路径指标和全路径指标作为路径组合指标的代表、SimRank 作为随机游走指标的代表，这些指标在各自类别中均具有优异表现[7]。通过对二分网络上的指标进行比较分析，发现不同路径对于合著关系预测的贡献大小和适合合著关系预测的最佳指标；通过比较二分网络上的指标与合著网络上的指标，发现路径权重对于合著关系预测的重要影响，以及影响合著关系预测的最重要影响因素。文献–作者二分网络和对应合著网络上的指标准确率和 AUC 值如表 2 和表 3 所示：

表 2　合著网络中合著关系预测指标的准确率和 AUC 值

指标	准确率|AUC (%)
共同邻居	27.1|85.4
全路径指标	25.5|86.5
局部路径指标 1	20.8|86.5
局部路径指标 2	25.5|86.5
资源分配指标	30.2|85.4
SimRank	12.9|85.7

表 3　文献–作者二分网络中合著关系预测指标的准确率和 AUC 值

二分网络路径指标	准确率|AUC (%)
二阶路径	48.3|85.5
三阶路径	28.6|86.0
路径组合(二阶路径+三阶路径)	59.1|86.6

合著关系预测的最佳指标是综合利用二阶路径和三阶路径信息的路径组合指标，表明不同长度的路径均对合著关系预测产生影响。在文献–作者二分网络和合著网络中，路径组合指标的准确率和 AUC 值均是最高的。其中，路径组合指标的准确率较二阶路径指标高出 10.8%，提高了 22.4%；较三阶路径指标高出 30.5%，提高了 63.1%；较合著网络中表现最好的资源分配指标高出 28.9%，提高了 95.7%；较合著网络中表现最差的 SimRank 指标高出 46.2%，提高了 3.58 倍。AUC 值的变化幅度不大，对指标进行宏观评测，不能对合著关系预测指标的优劣进行较好的区分。

路径组合指标针对特定数据集通过机器学习方法学习不同路径对于合著关系预测的影响，使得其成为合著关系预测的最佳指标，证实了权重的重要作用；同时，与合著网络的多个指标进行比较证实不同长度的路径对合著关系预测的作用并非一成不变，需要针对特定数据集进行学习和调整。在文献–作者二分网络和合著网络中，路径组合指标与局部路径指标均考虑了二阶路径和三阶路径的作用，不同的是，路径组合指标针对特定数据集学习了不同路径对合著关系预测的贡献，而局部路径指标则使用经验值(一般为 0.1 或 0.01)确定三阶路径相对于二阶路径的作用。表 2 和表 3 的准确率结果显示，路径组合指标较局部路径指标 1 高出 38.3%，提高了 1.84 倍；较局部路径指标 2 高出 33.6%，提高了 1.32 倍；较全路径指标高出 33.6%，提高了 1.32 倍。这些结果说明不同路径对合著关系预测具有不同作用，需要根据具体数据集进行针对性调整，从而得到最优结果，并且局部路径指标和全路径指标所采用的通用权重设置并不适合图书情报领域的合著关系预测，需要重新进行学习和调整。

二分网络上不同路径的权重证实二阶路径相对于三阶路径更重要，并且不同数据集上权重取值不同。如表 4 所示，在 10 折交叉验证构建的二分网络中，二阶路径和三阶路径的权重数值均不同，说明二阶路径和三阶路径对于合著关系预测的贡献需要针对特定数据集进行学习，并没有适用于多个数据集的最佳经验值。与此同时，表 4 中二阶路径的权重系数明显高于三阶路径的权重系数，说明二阶路径对于合著关系预测的贡献大大高于三阶路径；并且路径组合指标较二阶路径指标仅高出 10.8%，提高了 22.4%，证实共同邻居仍然是影响合著关系发生的最重要影响因素。

表 4　不同数据集构建的二分网络中不同路径的权重系数

数据集	二阶路径	三阶路径
1	2.97273259	–0.05000465
2	2.85770352	–0.0449394
3	2.58017868	–0.0439814
4	2.69195677	–0.04238217
5	2.18140025	0.02673774
6	2.97424309	–0.04686551
7	2.73535841	–0.04429512
8	2.79137504	0.00631618
9	2.46963496	–0.03885842
10	3.16311555	–0.04977438

合著网络上的预测指标同样证实共同邻居是合著关系预测的最重要影响因素。在表 2 中，局部路径指标和全路径指标均比共同邻居指标的准确率低，说明三阶或更高阶路径对于合著关系预测的贡献有限，并且使用固定的经验值作为多阶路径的权重时，这些多阶路径甚至会对合著关系预测产生负面影响。值得注意的是，资源分配指标作为共同邻居的直接改进指标，使用共同邻居的度区分作者对于合著关系预测的影响，它的准确率反而较共同邻居高出 3.1%，间接证实了二阶路径对于合著关系预测的重要影响。正因为如此，仍有大量研究从不同角度改进共同邻居指标，并取得了较好的效果[14-15]。

3.4　合著关系预测实例

本文选取合著网络中的最佳预测指标资源分配指标和二分网络中的两个最佳指标(二阶路径指标和路径组合指标)进行合著关系预测实例说明，并列出排名前 10 的合著作者对，如表 5 所示。其中，黑色斜体字表示预测成功的合著关系，未着重标出的为预测失败的合著关系。由于实验采用的是 10 折交叉验证，所以仅选取第一次实验结果进行说明，对应的指标正确率列在指标之后，如“资源分配指标(31.9%)”表示资源分配指标第一次实验的正确率为 31.9%。

表 5　三种指标预测出的排名前 10 合著关系

排名	资源分配指标(31.9%)		二阶路径指标 (49.2%)		路径组合指标(60.8%)	
1	***Detmer DE***	***Steen EB***	***Huntington P***	***Jamali HR***	***Zubair M***	***Jayakanth F***
2	Wang Y	Zhang L	***Nicholas D***	***Rowlands I***	***Poon EG***	***Jenter CA***
3	***Van Leeuwen TN***	***Costas R***	***Jamali HR***	***Rowlands I***	***Li JX***	***Zhang Z***
4	***Teo HH***	***Wei KK***	***Huntington P***	***Williams P***	***Chen YC***	***Hwang SJ***
5	***Nicholas D***	***Rowlands I***	***Bates DW***	***Jenter CA***	***Sia CL***	***Benbasat I***
6	***Narus SP***	***Evans RS***	***Markpin T***	***Sombatsompop N***	***Narus SP***	***Evans RS***
7	Bakken S	Lai AM	***Premkamolnetr N***	***Markpin T***	***Kaushal R***	***Lo HG***
8	***Accomazzi A***	***Kurtz MJ***	***Janssens F***	***Thijs B***	***Pan B***	***Lorigo L***
9	***Bates DW***	***Glaser J***	***Lee JH***	***Kang IS***	***Detmer DE***	***Steen EB***
10	***Huff SM***	***Staes CJ***	***Fox EA***	***Vemuri NS***	***Shea S***	***Cimino JJ***

由表 5 可看出，三种指标的预测效果均较好，其中资源分配指标成功预测出 8 对合著关系，而二阶路径指标和路径组合指标均成功预测出所有 10 对合著关系，说明路径组合指标是合著关系预测的最佳指标。与此同时，在排名前 20 和 30 的合著关系预测实例中，资源分配指标分别成功预测出 10 对和 14 对合著关系；二阶路径指标分别成功预测出 19 对和 27 对合著关系；路径组合指标分别成功预测出 19 对和 29 对合著关系。该结果再次证实投影为合著网络过程中的信息丢失对合著关系预测的负面影响以及二分网络上的路径相关指标能够更好地进行合著关系预测。

4　总结和展望

文献–作者二分网络在投影为合著网络过程中存在信息丢失，需要直接在二分网络上形成适合合著关系预测的指标和方法，并对合著关系形成的原因进行更好的分析和揭示。因此，本文在文献–作者二分网络上提出了一种基于路径组合的合著关系预测指标和方法，以提高合著关系预测的准确率和合著关系的可解释性。在图书情报领域的实验证实，二分网络上的二阶路径指标准确率明显高于合著网络上的共同邻居指标，并通过二者的准确率差异定量表示了二分网络投影为合著网络过程中的信息丢失，说明合著关系预测应直接在文献–作者二分网络上进行，以提高预测准确率和结果可解释性。另一方面，综合利用二阶路径和三阶路径信息的路径组合指标大大优于其他指标，说明不同路径均对合著关系预测产生贡献，但贡献程度需要针对特定数据集进行学习，而不能以通用的经验值进行指定；同时，二阶路径对合著关系预测的贡献明显高于三阶路径，说明了共同邻居仍是合著关系预测的最重要影响因素。

在图书情报领域的实验证实了利用路径组合指标进行合著关系预测的有效性，但还存在很多问题需要进一步研究，首先，四阶路径以及更多阶路径对于合

著关系预测的贡献还需进一步明晰，如在二分网络中提取出所有长度的路径并使用逻辑回归方法学习每种路径的权重系数，并以此形成全路径组合指标，对其准确率进行计算和比较。其次，基于逻辑回归构建路径组合指标的方法还需在其他领域进行实验，从而验证该方法的通用性，其他的机器学习方法也可引入到该模型中进行比较。最后，该方法可以扩展应用到其他类型的二分网络中，如专利–发明人二分网络上的发明人合作关系预测、微博–用户二分网络上的用户推荐和用户–商品二分网络上的商品推荐等。

参考文献：

[1] Barabasi A L, Jeong H, Neda Z, et al. Evolution of the Social Network of Scientific Collaborations [J]. Physica A: Statistical Mechanics and Its Applications, 2002, 311(3): 590-614.

[2] Guns R, Rousseau R. Recommending Research Collaborations Using Link Prediction and Random Forest Classifiers [J]. Scientometrics, 2014, 101(2): 1461-1473.

[3] Zhang Q, Xu X, Zhu Y, et al. Measuring Multiple Evolution Mechanisms of Complex Networks [J]. Scientific Reports, 2015, 5: Ariticle No. 10350.

[4] 张斌，马费成. 科学知识网络中的链路预测研究述评[J]. 中国图书馆学报，2015, 41(3): 99-113. (Zhang Bin, Ma Feicheng. A Review on Link Prediction of Scientific Knowledge Network [J]. Journal of Library Science in China, 2015, 41(3): 99-113.)

[5] Zhang J, Han T, Wang X. Uncovering the Mechanism of Knowledge Network Evolution by Link Prediction [J]. Geomatics and Information Science of Wuhan University, 2015, 39(S1): 100-106.

[6] Zhao J, Miao L, Yang J, et al. Prediction of Links and Weights in Networks by Reliable Routes [J]. Scientific Reports, 2015, 5: Ariticle No. 12261.

[7] Lv L, Zhou T. Link Prediction in Complex Networks: A Survey [J]. Physica A: Statistical Mechanics and Its Applications, 2010, 390(6): 1150-1170.

[8] Guns R. Bipartite Networks for Link Prediction: Can They Improve Prediction Performance?[C]. In: Proceedings of International Society for Scientometrics and Informetrics. 2011: 249-260.

[9] Gao M, Chen L, Xu Y. Projection Based Algorithm for Link Prediction in Bipartite Network[J]. Computer Science, 2016, 43(2): 118.

[10] 吴亚晶，张鹏，狄增如，等. 二分网络研究[J]. 复杂系统与复杂性科学，2010, 7(1): 1-12. (Wu Yajing, Zhang Peng, Di Zengru, et al. Study on Bipartite Networks[J]. Complex Systems and Complexity Science, 2010, 7(1): 1-12.)

[11] Daminelli S, Thomas J M, Duran C, et al. Common Neighbours and the Local-Community-Paradigm for Topological Link Prediction in Bipartite Networks[J]. New Journal of Physics, 2015, 17. http://iopscience.iop.org/article/10.1088/1367-2630/17/11/113037/meta.

[12] Zhou T, Lv L, Zhang Y C. Predicting Missing Links via Local Information[J]. The European Physical Journal B-Condensed Matter and Complex Systems, 2009, 71(4): 623-630.

[13] Hosmer Jr D W, Lemeshow S. Applied Logistic Regression [M]. New York: John Wiley & Sons, 2004.

[14] Güneş İ, Gündüz-Öğüdücü Ş, Çataltepe Z. Link Prediction Using Time Series of Neighborhood-Based Node Similarity Scores [J]. Data Mining and Knowledge Discovery, 2016, 30(1): 147-180.

[15] Sett N, Singh S R, Nandi S. Influence of Edge Weight on Node Proximity Based Link Prediction Methods: An Empirical Analysis[J]. Neurocomputing, 2016, 172: 71-83.

作者贡献声明:

张金柱，王小梅，韩涛：提出研究思路，设计研究方案，论文最终版本修订；
张金柱：采集、清洗和分析数据，完成实验，起草论文。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储，E-mail: zhangjinzhu@njust.edu.cn。
[1] 张金柱. wos_origin.rar. 从 Web of Sciences 上下载的原始数据.
[2] 张金柱. data2.rar. 处理后的文献–作者数据.

收稿日期: 2016-06-15
收修改稿日期: 2016-08-01

Predicting Co-authorship with Combination of Paths in Paper-author Bipartite Networks

Zhang Jinzhu[1] Wang Xiaomei[2] Han Tao[2]
[1](School of Economics and Management, Nanjing University of Science and Technology, Nanjing 210094, China)
[2](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: [**Objective**] This paper aims to predict co-authorship more effectively and reduce the information loss. [**Methods**] First, we constructed a paper-author bipartite network and its co-authorship counterpart in the field of library and information science. Second, we described the relationships among authors with the path-length of two and three from the bipartite network. Third, we used the logistic regression method to learn the influence of different factors. Finally, we predicted co-authorship in the paper-author bipartite network with various indictors. [**Results**] We found significant information loss in the change from the paper-author bipartite network to the co-authorship network. The logistic regression method was an appropriate way to learn the contributions of paths. The new indicators were more accurate and the predicted co-authorships could be interpreted more easily. [**Limitations**] We did not include the multiple paths methods to the present study and more research is needed to examine the proposed method in other areas. [**Conclusions**] Co-authorship prediction should be conducted in the paper-author bipartite network to reduce the information loss. The paths combination indicator in the paper-author bipartite network might be the most effective method to predict co-authorship, which could be applied to the patent-inventor bipartite network.

Keywords: Paper-author bipartite network Paths combination indicator Library and Information Science Co-authorship network Co-authorship prediction

美国图书馆和信息资源委员会获得 270 万美元的项目资助，以保护面临风险的数据记录

Andrew W. Mellon 基金会向美国图书馆和信息资源委员会(CLIR)提供了高达 2,725,000 美元的项目资助，用于重新计划将具有高学术价值的、“有风险”的音像材料进行数字化。该项目将在 2017 年 1 月至 2018 年 9 月期间举办 4 次比赛，奖金总额高达 230 万美元。

为制定新的指导方针和标准，CLIR 将于 2017 年 1 月与 NEDCC 合作发布一项试点呼吁以寻求建议。试点呼吁将仅集中于磁带音频媒体的重新格式化，通过 NEDCC 的扩展音频保存服务进行数字化。CLIR 将召集一个独立审查小组进行评估。经审核后，CLIR 将支付总额高达 150,000 美元，每项资助从 5,000 美元到 25,000 美元不等，直接用于支付 NEDCC 提供的音频重新格式化服务的费用。

之后，CLIR 将发起一系列共三个公开竞赛，预计在两年内发放 215 万美元的资金。三项公开赛的征集将分别于 2017 年 6 月、2017 年 12 月和 2018 年 5 月发出。公开比赛的奖金将在 1 万美元至 5 万美元之间，包括音像和视听内容的重新格式化所涉及的直接费用。

(编译自: https://www.clir.org/about/news/pressrelease/recordings-at-risk)

(本刊讯)